

Analytics Drives Big Data Drives Infrastructure

Confessions of Storage turned Analytics Geeks

Dr. Alope Guha

29th IEEE Conference on Massive Data Storage

May 8th, 2013

aloke@cruxly.com

What's Common Between
a Sensor that could Distinguish a fine Cognac,
and Predicting Movies You'd Like on Netflix?

The Sommelier "Robot"

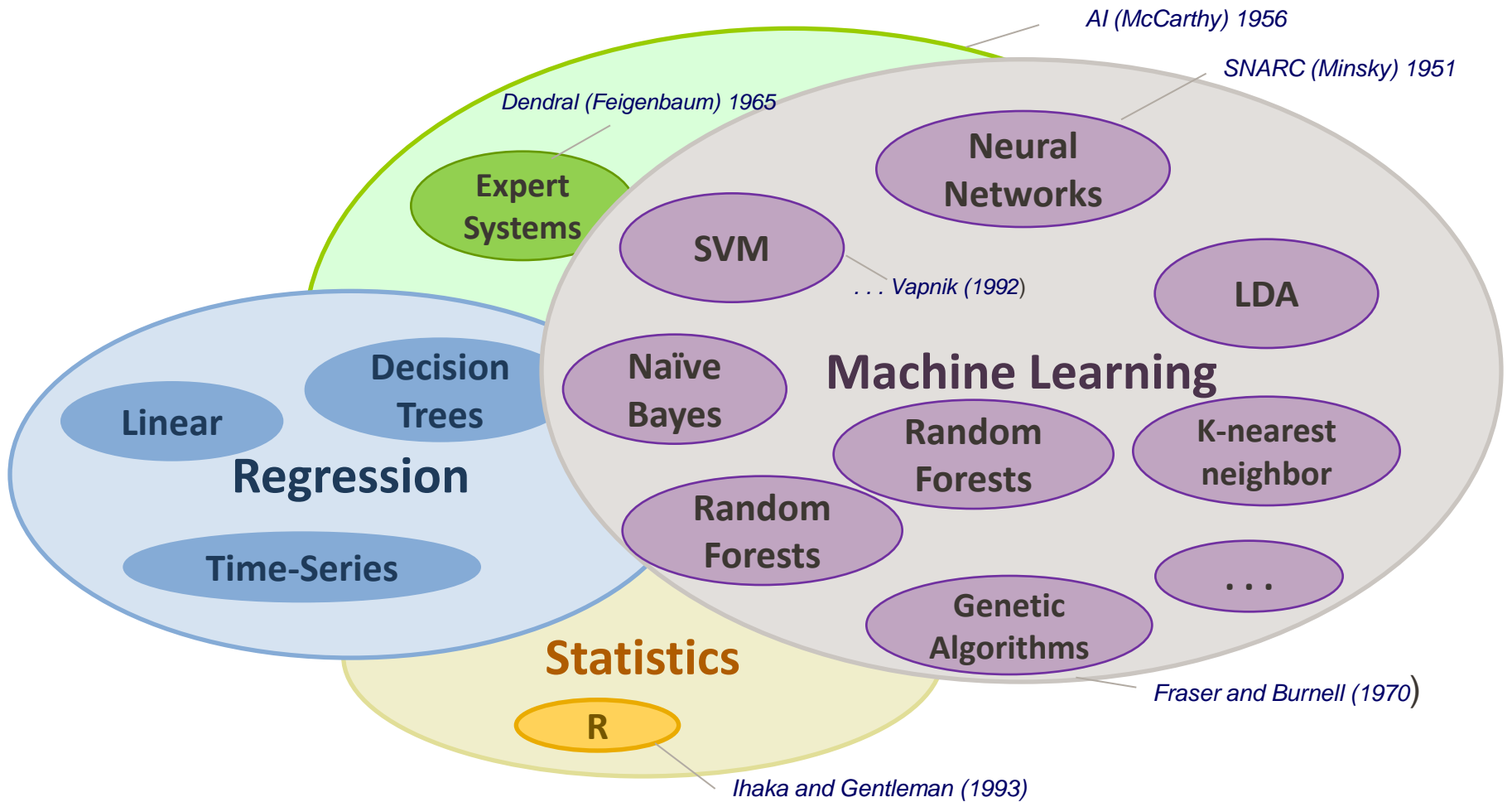


Predicting What Movies You'd Watch



(Analytics, BigData, DataStore)+

Many Analytics Techniques . . .



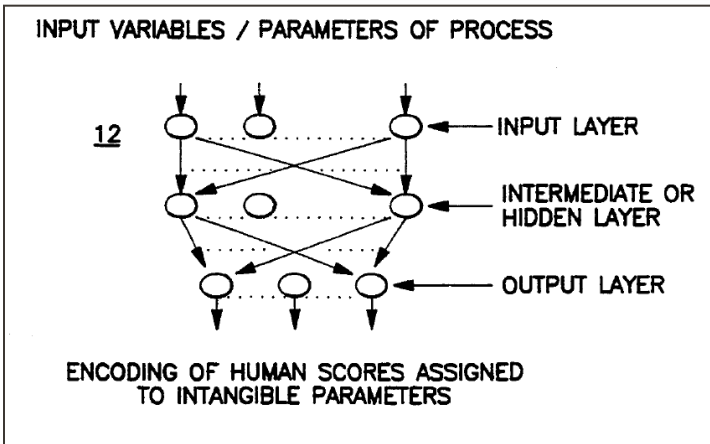
Common Analytics Processing pre-2000

- Sources: Local
- Data: Numeric, Homogeneous
- Processing: Local
- Consumer: Local

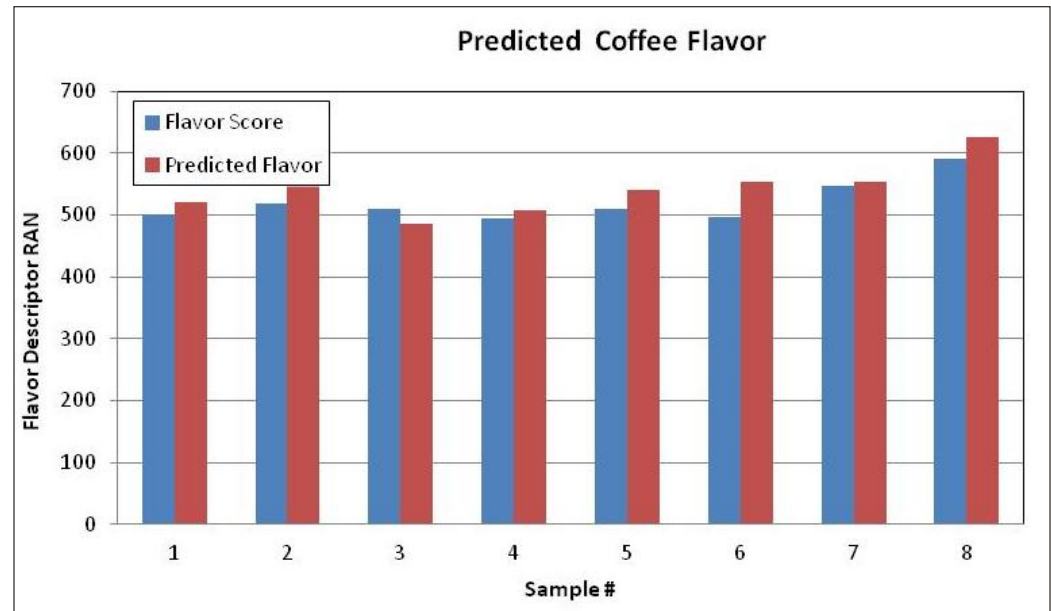
- Analytics: Linear/Non-Linear Regression, Neural Networks, SVM, LDA, LSA, Decision Trees, Monte Carlo, Lin-Ops, Expert Systems . . .



Flavor Predictor – Neural Networks

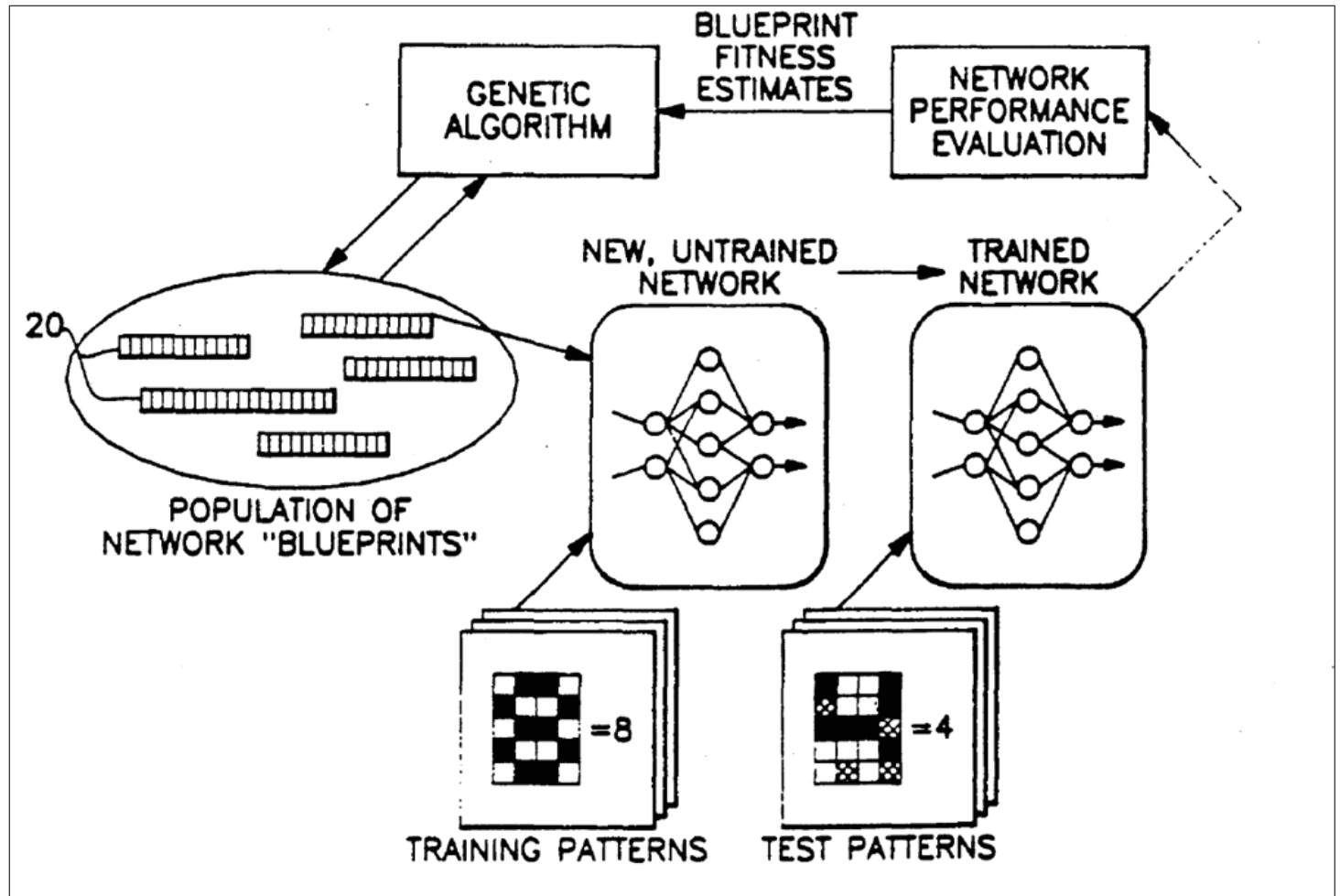


USPTO #5,373,452 (1994)



1988

Pattern Recognition – Genetic Algorithms



US PTO #5,140,530, 1992

Small to Big

The Meaning of Big Data - 3 V's

- Big Volume
 - Simple (SQL) analytics: **Data Warehouses**
 - Complex (non-SQL) analytics: **emerging market**
- Big Velocity
 - Drink from the fire hose:
complex event processing, NoSQL, New SQL
- Big Variety
 - Large number of diverse data sources to integrate:
data integration, ETL

Typical Analytics: 2000-2006

- Sources: Global , Social Networks
- Data: Heterogeneous, Numeric, Text
- Processing: Hosted/Scale
- Consumer: Global

- Analytics: Batch Mode, Social Media Marketing, Churn Detection, Sentiment Analysis, etc.



2007- : Internet Data Analytics

YAHOO!

Google™

CNN

YouTube

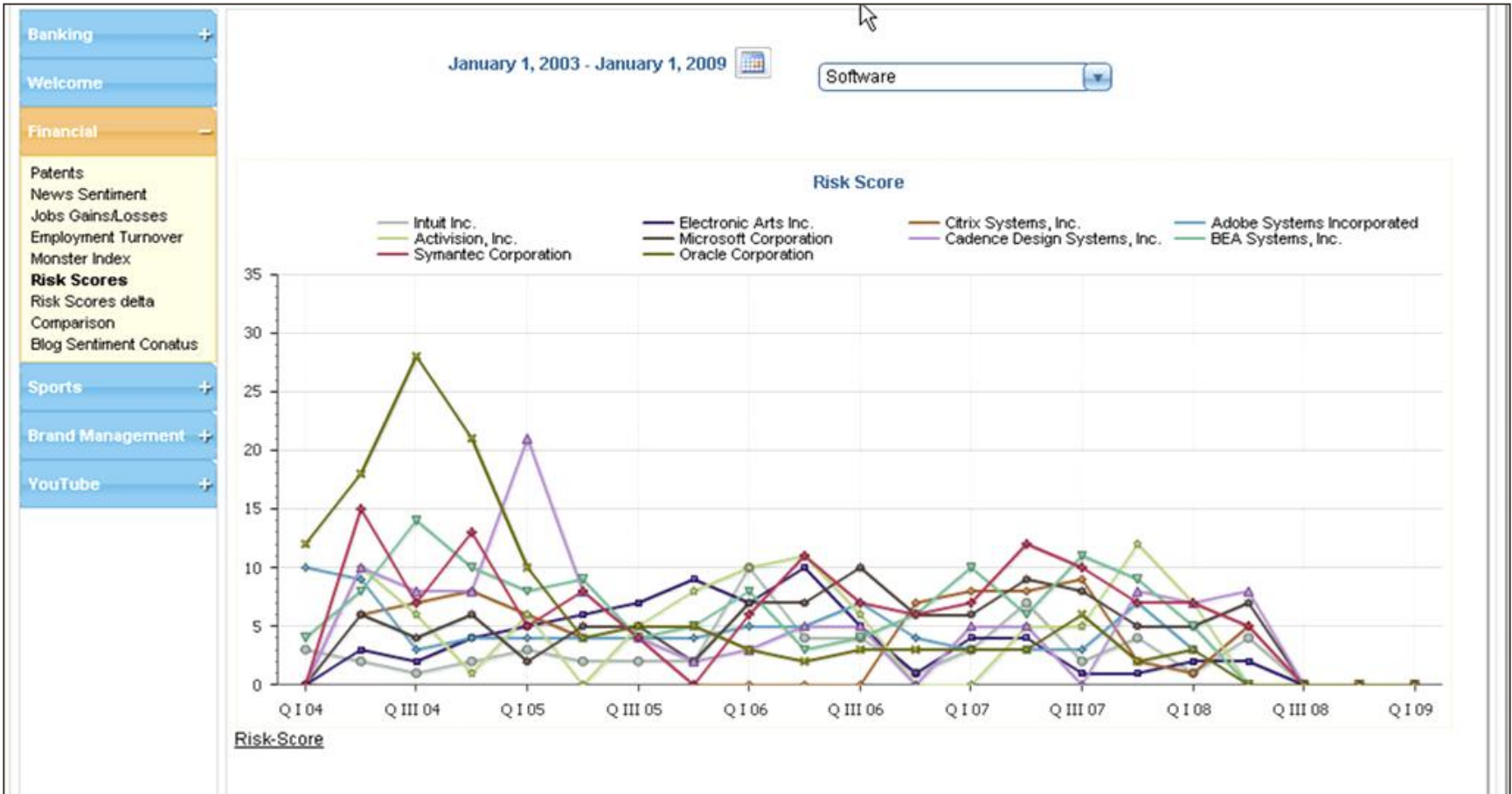


Analytic	Description
Compliance	<input type="checkbox"/> Detect customer-defined violations, governance <input type="checkbox"/> Detection of specified events: e-Discovery
IP Leakage	<input type="checkbox"/> Email/IM <input type="checkbox"/> Voice calls (ongoing)
Detect Improper Content	<input type="checkbox"/> Custom definitions: racy, harassment, etc..
Summary of Content	<input type="checkbox"/> Derive topic-specific summaries at sub-document level: public or corporate archival data
Share of Voice	<input type="checkbox"/> # of mentions, # of features, audience count <input type="checkbox"/> By publications, by relative coverage vs. competitors
Opinion/Sentiment	<input type="checkbox"/> Customer comments: blogs, phone calls, web (CRM) <input type="checkbox"/> Spokesperson effectiveness

*Aumni*data

Financial Risk Scoring: Detect

*Aumni*data



Risk Scoring: detect incremental change in # occurrences where corporate officers mention “risk” (or equivalent terms) during earnings call

Financial Risk Scoring: Listen

*Aumni*data

Banking Analytics

- Net ChargeOff: 4 Banks
- ChargeOff Ratio: 4 Banks
- Net ChargeOff: 2 Banks
- ChargeOff Ratio: 2 Banks
- Net ChargeOff: 1 Bank
- Past Due Loans
- ChargeOff Ratio <90 days
- ChargeOff Ratio >90 days
- ChargeOff Ratio
- Nonaccrual
- Sentiment

Financial 2

- Patents
- Sentiment
- Jobs Gains/Losses
- Employment Turnover
- Monster Index
- Risk Scores
- Risk Scores delta Comparison**

Conatus Blogs

- Sentiment

Soft Drinks

- Sentiment
- Positive Mentions
- Negative Mentions

My Portfolio

January 1, 2005 - January 1, 2009

Windows Media Player

Paused 01:13

Risc-Score delta

Legend: Google Inc. (black), Electronic Arts Inc. (orange), Amazon Inc. (teal), Celgene Corporation (green)

Callout for Celgene Corporation: 8, 01 April 2008 00:00:00 - 01 July 2008 00:00:00

DrillDown Options -- Webpage Dialog

Drill Down Series: Celgene Corporation

By Dimension

Drilldown to event 'Risk-Score' - Windows Internet Explorer

http://localhost/InsightViewer/OnLineData/KPIGrid2.aspx?gridId=&partId=&dataKey=30a2e1b9-8fa9-484b-b9fa-99eba9b09008

EventDate	company_name	risk_count	derived_audio_URL	derived_doc_URL	Relevance Class	Class_Company
05/08/2008 12:00 AM	Celgene Corporation	12.00	http://localhost/cp	http://localhost/cp-transcript-2008-05-08.txt	My Portfolio	My PortfolioCelgene Corporation

[Export to excel](#) | [Print](#)

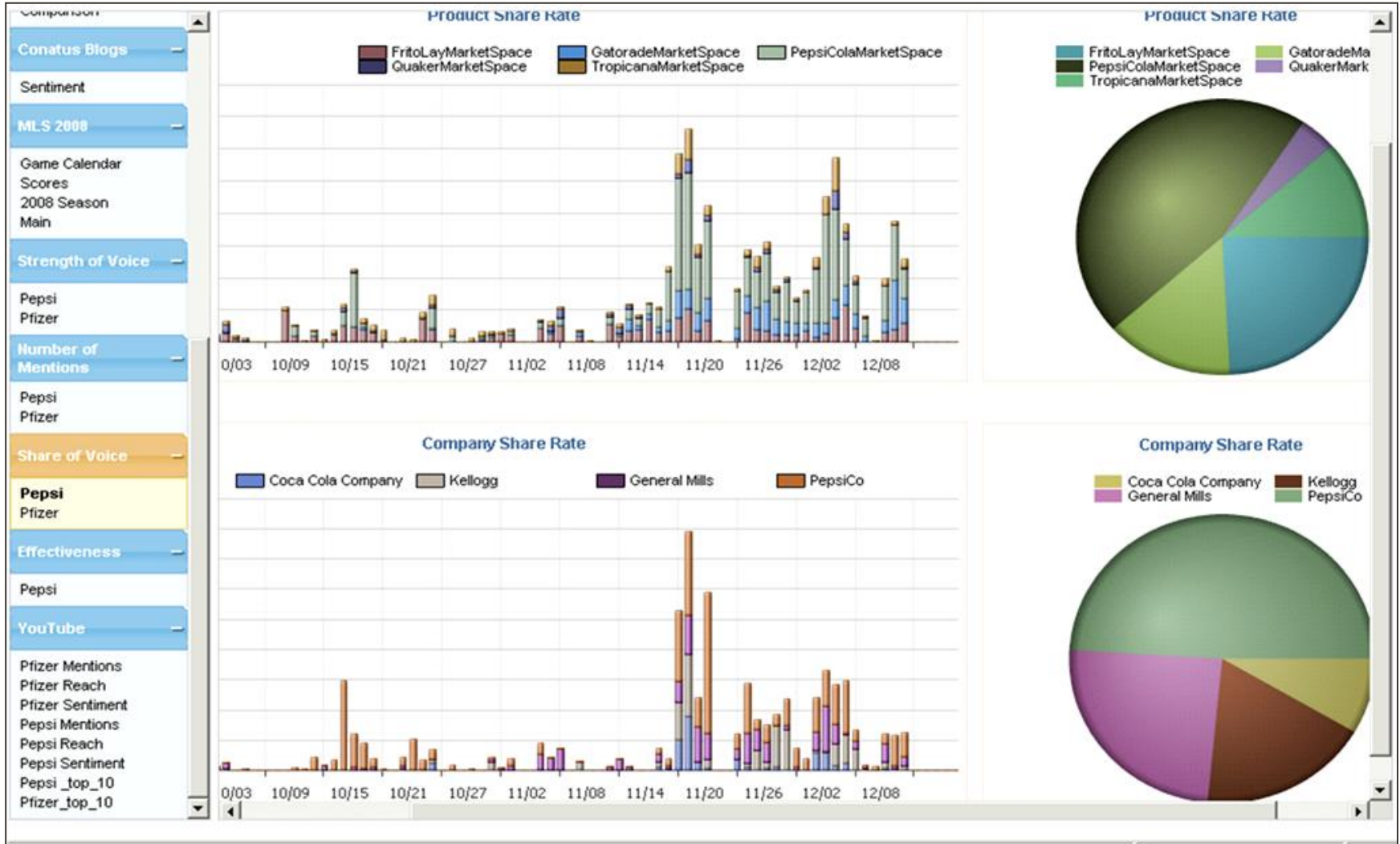
Banking: Credit Worthiness – remember 2008?



Analyze bank reports to assess loans, payments, recoveries, etc. for key bank indexes, groups of banks, or individual banks

Share of Voice: Online Buzz

*Aumni*data



Sentiment Analysis

*Aumni*data

The screenshot shows a Yahoo! Finance news article titled "Celgene shares rise on upgrade, market recovery". The article text includes: "Monday October 13, 8:51 pm ET", "Celgene shares gain ground on analyst upgrade ahead of earnings as broader market bounces back", and "NEW YORK (AP) -- Shares of Celgene stock market, as the biotechnology industry recovers from recent results and Robert W. Baird upgrade".

Below the article, an advertisement is visible. A second browser window, titled "Drilldown to event 'Sentiment'", displays a table of sentiment analysis results for the event.

EventDate	score	derived_doc_URL	company_name	Relevance Class
10/13/2008 12:00 AM	5.00	http://localhost/cp/web/News/10132008/Yahoo%20Finance%20Markets%20US%20Markets%20797.html	Celgene Corporation	My Portfolio

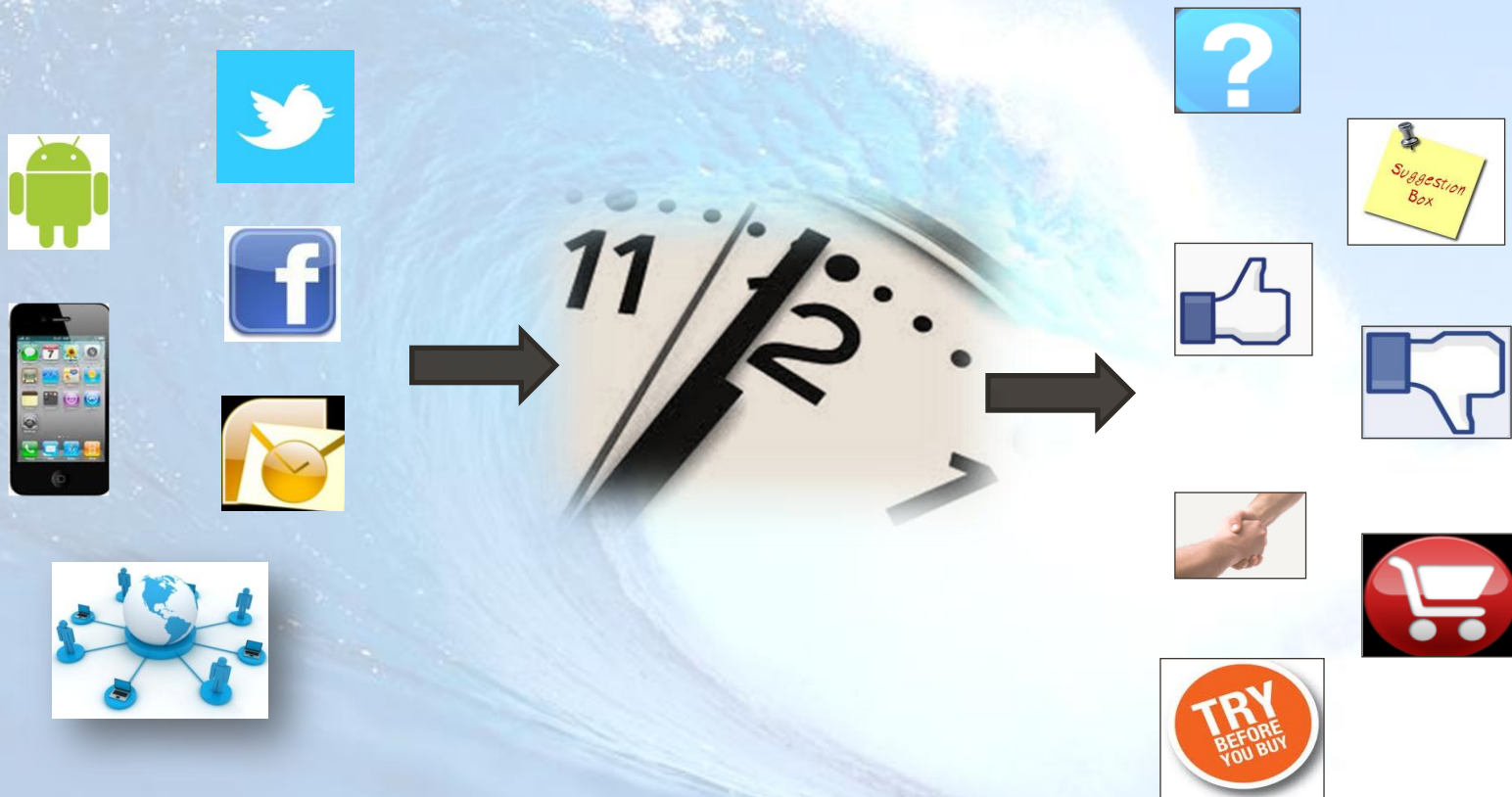
Additional elements in the screenshot include a line chart for Celgene Corporation (CELG) showing stock price fluctuations, and various navigation and search options on the Yahoo! Finance interface.

Analytics Processing: 2007-

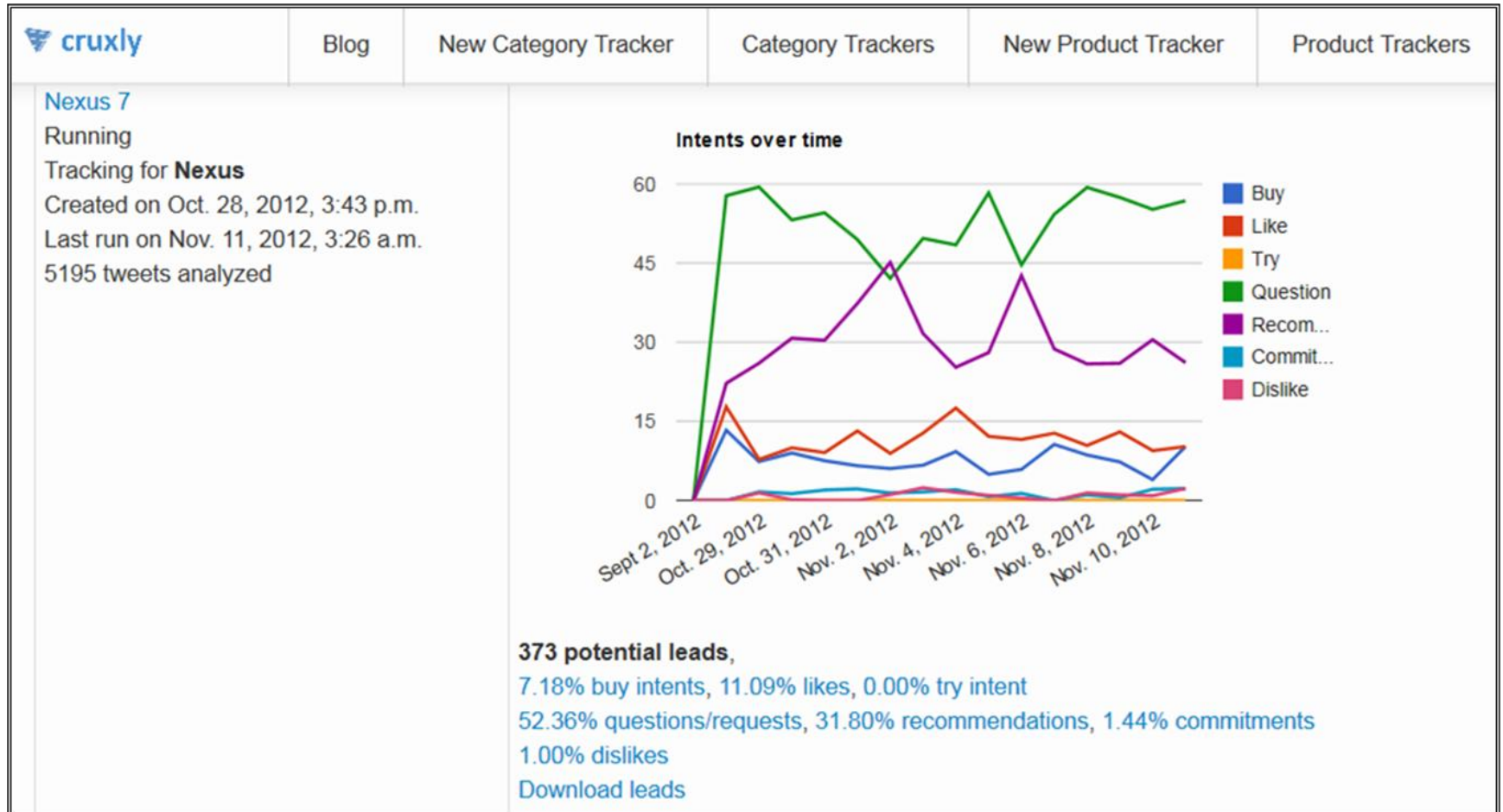
- Sources: Global, Mobile, New Social (Instagram, . . .)
 - Data: Multi-Dimensional, Heterogeneous, Audio/Video
 - Processing: Hosted/Scale
 - Consumer: Global
-
- Analytics: Batch, Streaming, . . .




2008 - : Real-Time/Streaming Analytics




Brand Marketing



Brand Management


 [Blog](#) [New Category Tracker](#) [Category Trackers](#) [New Product Tracker](#) [Product Trackers](#)

Like intents for Nexus 7




crapulent

Saw a Nexus 7 at work today. I like that place a lot.
2012-11-08 16:44 [Reply](#) [Retweet](#) [Favorite](#)




headdeskben

[@GeEki_chan](#) I've got the nexus 7 myself, and I love it :D
2012-11-08 16:29 [Reply](#) [Retweet](#) [Favorite](#)




pedalpusherpunk

[@liversedge](#) you are right. Separately I am loving the Nexus 7, have it for over a week now but seems longer. Only bad is lack of tablet apps
2012-11-08 16:14 [Reply](#) [Retweet](#) [Favorite](#)




VA5LF

So far I am liking this Nexus 7 a lot.
2012-11-08 14:59 [Reply](#) [Retweet](#) [Favorite](#)




Wtdarkandwild

[@unklerupert](#) [@SongWarmonger](#) Yep suited me as an Amazon gal but the Nexus 7 is lovely.
2012-11-08 14:14 [Reply](#) [Retweet](#) [Favorite](#)




[@AlexWilliams03](#) The Nexus 7 is really nice. Plus, you could get 2 of 'em for the price of one Nexus 10. O_o

Customer Support

	Blog	New Category Tracker	Category Trackers	New Product Tracker	Product Trackers	Logout admin
---	----------------------	--------------------------------------	-----------------------------------	-------------------------------------	----------------------------------	------------------------------


Dislike intents for jetblue



metschick

@hip_hip_jorge Aw, boo. I'll be flying Jetblue in 4 weeks to Orlando.


2012-10-24 18:25 [Reply](#) [Retweet](#) [Favorite](#)



FleetWavyAss

Highed up, thinkin bout the best move next.. Like fuck jetblue, i need a blue jet

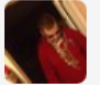
2012-10-24 18:25 [Reply](#) [Retweet](#) [Favorite](#)



xoNELLY

Stupid JetBlue!! Why the fuck you can't fly out of Miami Airport ☹


2012-10-24 18:25 [Reply](#) [Retweet](#) [Favorite](#)



daghernation

Fuck jetblue I need a blue jet HAAAN

2012-10-22 14:29 [Reply](#) [Retweet](#) [Favorite](#)



LDalia

Flight delayed 1 1/2 hrs. JetBlue sucks or maybe it's the Ft. Lauderdale Aport

2012-10-22 12:49 [Reply](#) [Retweet](#) [Favorite](#)

Customer Support


The image shows a screenshot of a web browser with two overlapping windows. The background window displays a Twitter thread on the 'cruxly' website. The thread includes tweets from 'FleetWavyAss', 'xoNELLY', 'daghernation', and 'LDalia'. The tweet from 'LDalia' is the focus of the foreground window.

The foreground window is a 'Reply to a Tweet' form for the tweet by Lauren @LDalia. The tweet text is: "Flight delayed 1 1/2 hrs. JetBlue sucks or maybe it's the Ft. Lauderdale Aport" (dated 10:45 AM Oct 22nd). The reply input field contains the text: "@LDalia | Sorry you experienced this delay. Because of hurricane Sandy, we could not leave per schedule." The 'Tweet' button shows 132 replies.


Background tweets (from top to bottom):

- FleetWavyAss**: Highed up, think... 2012-10-24 18:2...
- xoNELLY**: Stupid JetBlue!! 2012-10-24 18:2...
- daghernation**: Fuck jetblue I ne... 2012-10-22 14:2...
- LDalia**: Flight delayed 1 1/2 hrs. JetBlue sucks or maybe it's the Ft. Lauderdale Aport 2012-10-22 12:49 Reply Retweet Favorite
- @JetBlue**: I know they are. That's why it was so disheartening when they weren't today. I tried. They didn't.


Lead Generation

 Blog New Category Tracker Category Trackers New Product Tracker Product Trackers


Buy intents for samsung




I want one of those Smart TV's
2012-10-05 22:56 Reply Retweet Favorite




Finally, everyone gone for a weekend. Now I can really use that 51' smart tv for what it's ment
for # yesitiswhatyourthinking
2012-10-05 17:56 Reply Retweet Favorite



@Kerigilles @catelynnlowell IMA GET IT ONE DAY LOL it's not a the smart tv ☺
2012-10-05 12:00 Reply Retweet Favorite




something comes out on tv bro: OH I WANT THAT!!!!!! Me: oh my *gives death stare* NO I WANT IT FIRST haha he's 2 -_-
2012-10-05 11:07 Reply Retweet Favorite



I want the new smart tv for Christmas it's amazing 😊 # samsung.
2012-10-04 15:31 Reply Retweet Favorite

Opening cruxly-intents-samsung-2012_10_24_20_12_46_305682.xls

You have chosen to open:

 **cruxly-intents-samsung-2012_10_24_20_12_46_305682.xls**
which is a: Microsoft Office Excel 97-2003 Worksheet
from: <http://www.cruxly.com>

What should Firefox do with this file?

Open with Microsoft Office Excel (default)

Save File

Do this automatically for files like this from now on.

OK Cancel

... More Data, Faster

CIO INSIGHT.

Data Analytics Allows P&G to Turn on a Dime

By Peter High | Posted 05-03-2013

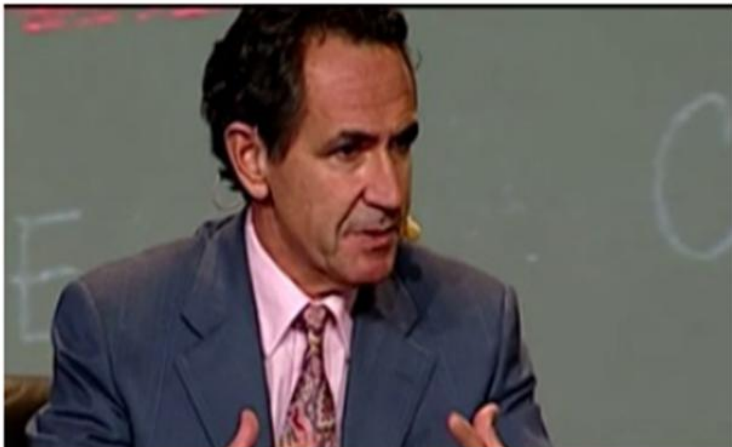


Print



Email

Filippo Passerini provides an overview of the steps that P&G took to improve its analytic capabilities and harness the power of big data in real-time.



By Peter High

Filippo Passerini, CIO and Group President of Global Business Services of Procter & Gamble, discusses the approach he and his team have taken to get better, more accurate data analysis into the right executives' hands in a timely fashion. The result is a remarkable track record of innovation.

IN SUMMARY

WHO: Filippo Passerini, CIO and Group President of Global Business Services of Procter & Gamble

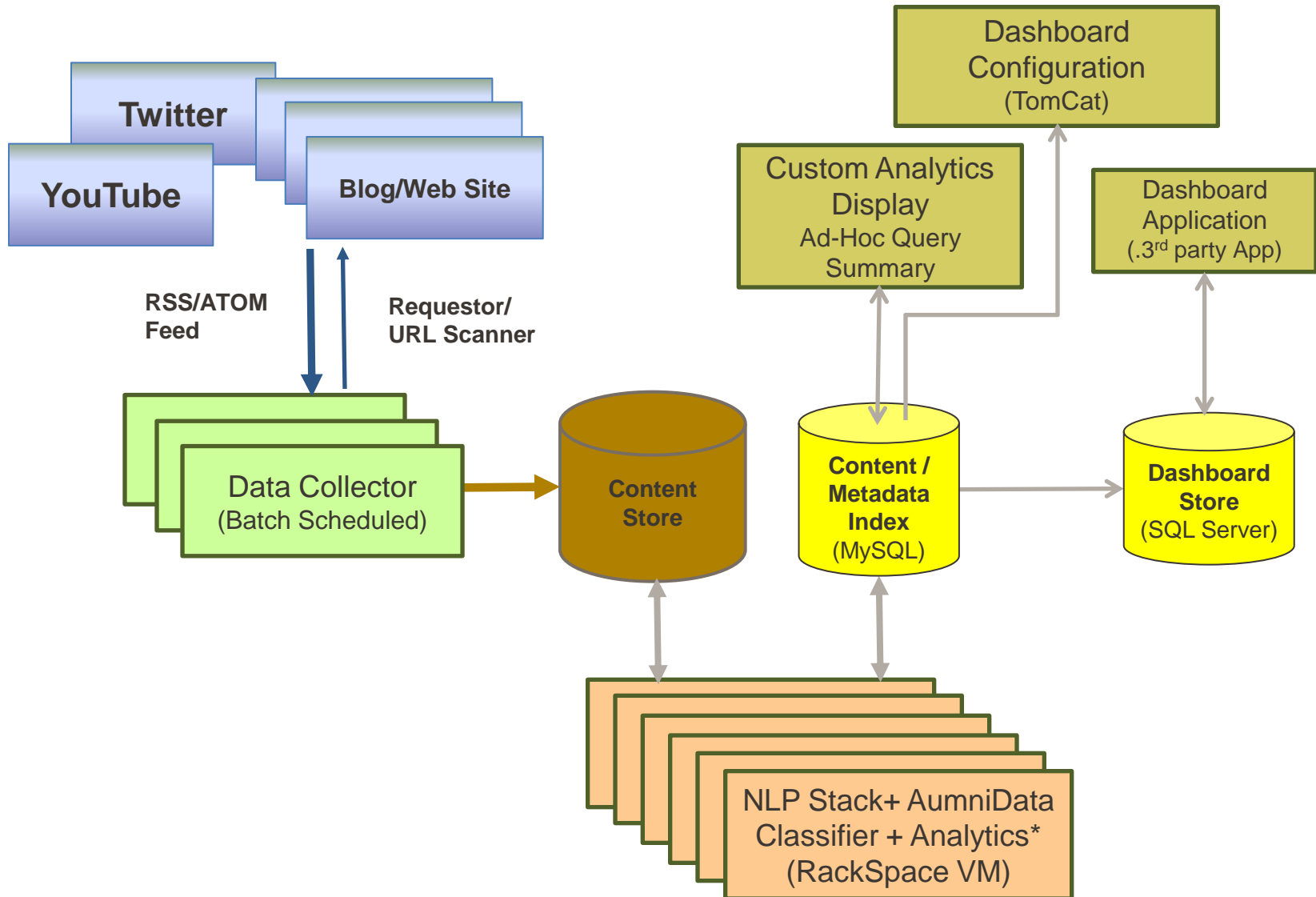
<http://www.cioinsight.com/it-strategy/big-data/data-analytics-allows-pg-to-turn-on-a-dime/?kc=CIOMINUTE05062013CIOA>

“Internet of Things”

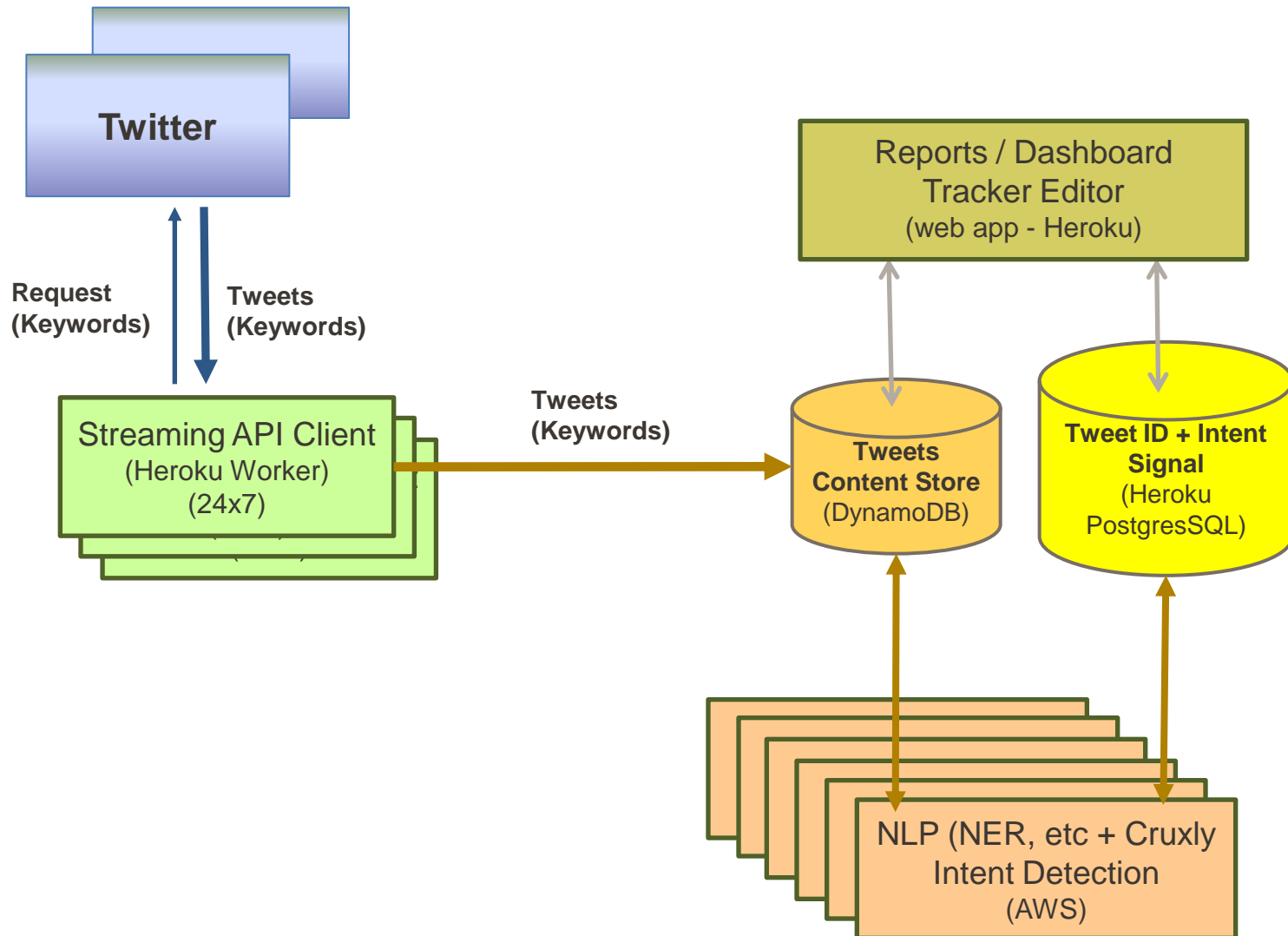


<http://www.news-sap.com/survey-by-sap-and-harris-interactive-finds-brazil-china-germany-and-india-most-ready-for-m2m-technology-to-drive-connected-smarter-cities/>

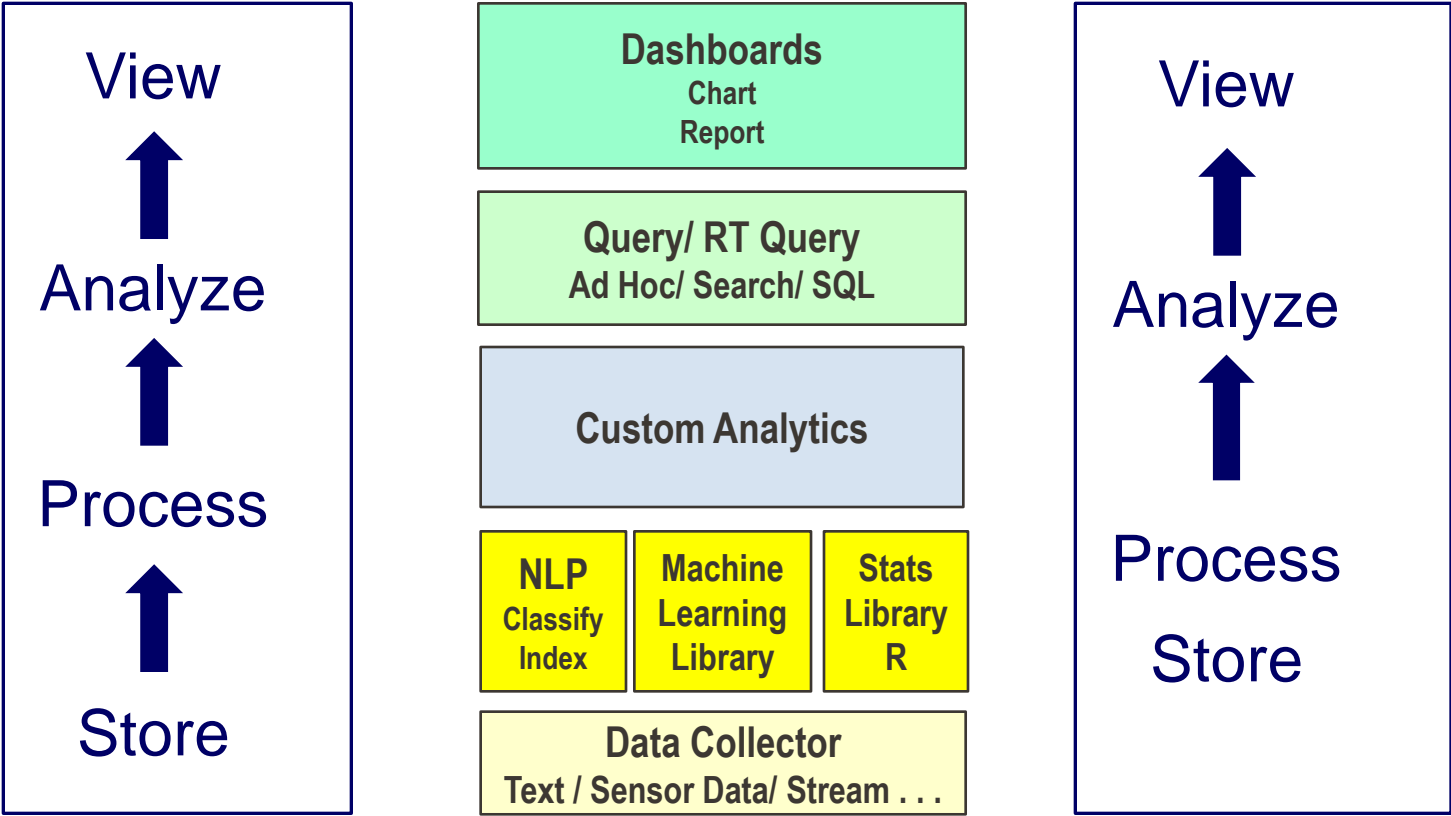
AumniData: Batch Processing



Cruxly: Stream Processing



Data Analytics Demands . . .



Storm

S4 distributed stream computing platform

Yarn



Storage Implications: Back to the Future



MB/s – Batch

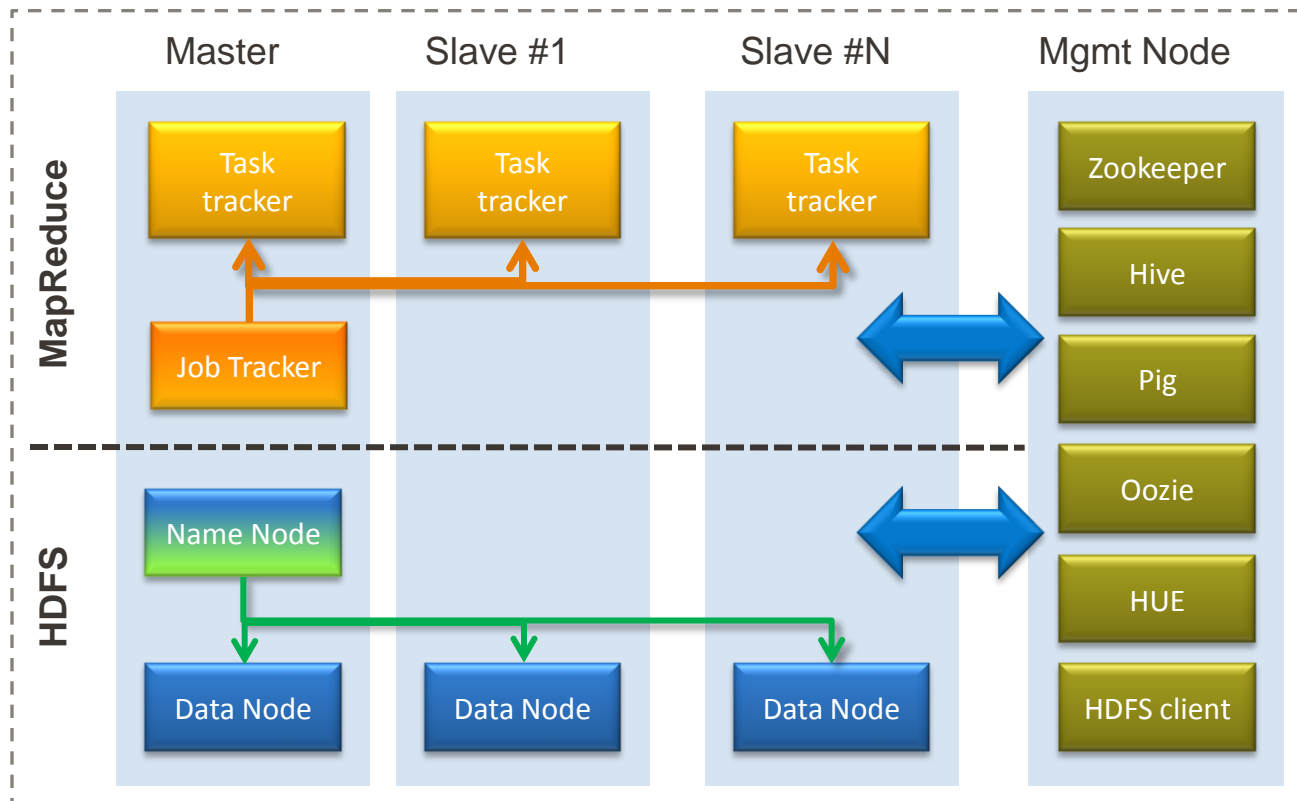


IOPs – Stream



Both?

Storage Implications: Back to the Future II, III

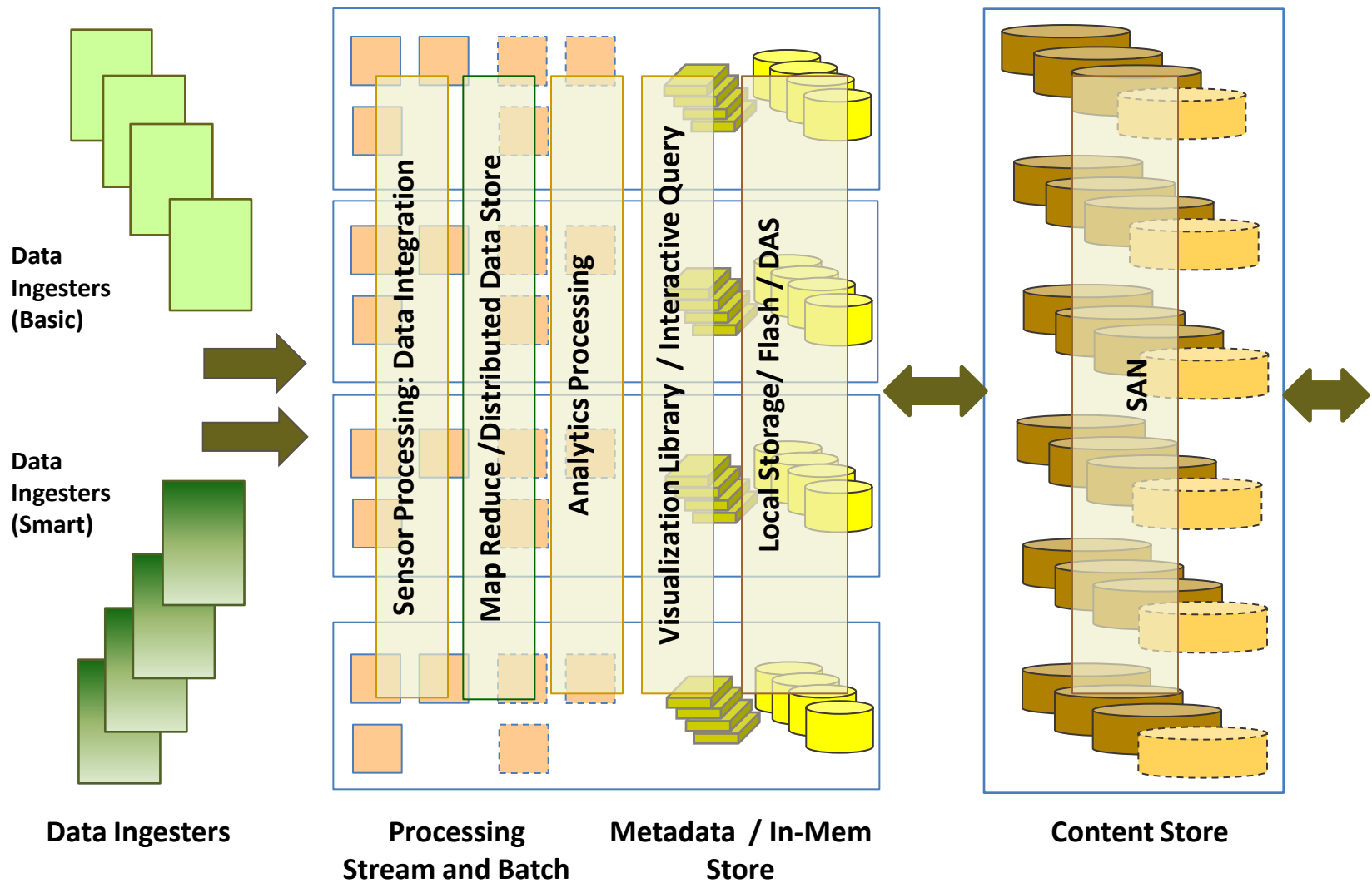


Storage Capacity Scaling?

Import/Export Data?

Storage Tiering?

A More General Data Analytics Framework?



Conclusion

- Data Analytics \Rightarrow Big Data \Rightarrow Scale-Out
- Variety \Rightarrow Infrastructure
- Volume \Rightarrow Bandwidth Support
- Velocity \Rightarrow Streaming Support
- We Solved the Processing Problem
- We Need to Solve the Larger Storage Problem



Grateful Acknowledgements

- Kapil Tundwal
- Dr. Kirill Kireyev
- Dr. Andrew Lampert
- Venky Madireddy
- Dr. Shumin Wu
- Joan Wrabetz